Uchaguzi-2022: A Dataset of Citizen Reports on the 2022 Kenyan Election

Roberto Mondini[†] **Neema Kotonya**[†] **Robert Logan**[†] **Elizabeth Olson**[†]
Angela Oduor Lungati[‡] **Daniel Duke Odongo**[‡] **Tim Ombasa**[‡]
Hemank Lamba[†] **Aoife Cahill**[†] **Joel Tetreault**[†] **Alejandro Jaimes**[†]
[†]Dataminr Inc. [‡]Ushahidi
{rmondini,neema.kotonya,rlogan,elizabeth.olson,
hlamba,acahill,jtetreault,ajaimes}@dataminr.com
{angela,daniel,tim.ombasa}@ushahidi.com

Abstract

Online reporting platforms have enabled citizens around the world to collectively share their opinions and report in real time on events impacting their local communities. Systematically organizing (e.g., categorizing by attributes) and geotagging large amounts of crowdsourced information is crucial to ensuring that accurate and meaningful insights can be drawn from this data and used by policy makers to bring about positive change. These tasks, however, typically require extensive manual annotation efforts. In this paper we present Uchaguzi-2022, a dataset of 14k categorized and geotagged citizen reports related to the 2022 Kenyan General Election containing mentions of election-related issues such as official misconduct, vote count irregularities, and acts of violence. We use this dataset to investigate whether language models can assist in scalably categorizing and geotagging reports, thus highlighting its potential application in the AI for Social Good space.[1]

1 Introduction

Citizen journalism (i.e., non-professional reporting disseminated on social media and dedicated websites) plays an increasingly important role in enabling the circulation of opinions on topics such as elections, as well as in exposing electoral violence and interference (Moyo, 2009; Ajao, 2017). To ensure credibility, it is incumbent on social media platforms and websites to verify citizen-submitted reports, since unmoderated content can exacerbate the spread of misinformation (Ndlela, 2020). As reports are often issued via social media posts, it is also crucial for platforms to be able to *categorize* (e.g., by topic) and *geotag* (i.e., add geographic metadata to) posts to help readers understand how the reported events are impacting their communities. Furthermore, the systematic organization of large amounts of crowdsourced information enables reporting agencies to draw meaningful insights from this data and share these with the relevant policy makers to enact positive changes (Shayo, 2021). Unfortunately, categorizing and geotagging large amounts of data typically requires substantial manual effort, often led by volunteers (Aarvik, 2015; Shayo, 2021), severely delaying and limiting the publication and circulation of information.

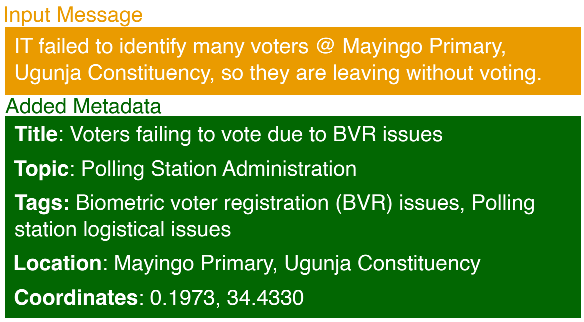


Figure 1: **Annotated report example**. Each input message *(top)* is categorized and geolocated *(bottom)*.

In this paper we introduce Uchaguzi-2022, a dataset of 14k reports pertaining to the 2022 Kenyan General Election that have been categorized and geotagged (Figure 1). Unlike existing social media-based election datasets, e.g., Wang et al. (2012), Schmidt et al. (2022), that focus on studying electorate sentiment and community structure, Uchaguzi-2022 contains reports of election interference issues such as official misconduct, vote count irregularities, and acts of violence. Through this dataset we explore NLP methods to automate report categorization and geotagging. We provide benchmark results for both encoder-only language models (Conneau et al., 2020) trained using fine-tuning (Devlin et al., 2019), and decoder-only language models (Jiang et al., 2024) trained using few-shot in-context learning (Brown et al., 2020). By al-

[1]Dataset and code available at: `https://github.ushahidi.org/uchaguzi-ai/`

Proceedings of the 31st International Conference on Computational Linguistics, pages 5807–5825
January 19–24, 2025.

leviating the manual efforts to complete these tasks, systems trained on this dataset will allow reporting organizations to focus on extracting meaningful insights from this structured data and help inform decisions that can positively impact the reporting citizens and their communities.

In summary, our contributions are as follows:

1. We present Uchaguzi-2022, a dataset of categorized and geotagged citizen reports on the 2022 Kenyan General Election. To our knowledge, this is the first dataset of citizen-contributed election issues in the African continent that contains this associated metadata.
2. We use this dataset to benchmark models for assisting with categorization and geotagging, demonstrating the potential application of this data for broader AI for Social Good efforts. Our results show that few-shot models are competitive with fully fine-tuned models.

2 The Uchaguzi-2022 Dataset

2.1 Overview

Uchaguzi-2022 contains 14,169 citizen reports related to the Kenyan General Elections held on August 9, 2022 and submitted to the Uchaguzi platform[2] between June 27 and August 29, 2022. This platform, developed and maintained by Ushahidi[3], enables citizens to share their views and report on election-related events through SMS messages, X (formerly Twitter) posts, and questionnaires initiated via USSD and WhatsApp. The submitted reports are manually reviewed and annotated by a team of Ushahidi volunteers, who additionally can create their own reports from public social media posts. Volunteers are fluent in both English and Swahili and participate in training sessions prior to undertaking the annotation tasks (Ushahidi, 2022b). After the volunteers' review, reports are surfaced to the public and reporting agencies through the Uchaguzi website (Ushahidi, 2022a).

Volunteers *categorize* incoming reports in two ways: 1) by assigning a *topic*, and 2) by assigning topic-specific *tags*. There are ten topics (Table 1), while tags provide further fine-grained categorization. The set of applicable tags is dependent on the topic selected, is optional, and multiple tags can be assigned to a report. For illustration, the tag set for the *Polling Station Administration* topic is listed in Table 2, and the tag sets for the remaining topics

[2]https://uchaguzi.or.ke/

[3]https://www.ushahidi.com/

Topic	Count
Opinions	9,441
Media Reports	1,736
Positive Events	1,256
Counting and Results	755
Security Issues	302
Voting Issues	285
Political Rallies	134
Polling Station Administration	123
Staffing Issues	76
Irrelevant Report	61

Table 1: **List of topics and their prevalence**.

Tag	Count
Biometric voter registration (BVR) issues	38
Polling station logistical issues	38
Low voter turn out	23
Missing/inadequate voting materials	14
Ballot box irregularities	9
Polling station closed before voting concluded	5
Campaign material in polling station	3
High voter turn out	2

Table 2: **Tags for the *Polling Station Administration* topic and their prevalence**.

as well as examples for each topic are provided in Appendix A.

As tags are added, volunteers also give reports a *title*, i.e., a short summary of the content of the report, and *geotags*. Geotags are provided in the form of *coordinates* (latitude and longitude) as well as an optional *location* name, and are determined by looking for location mentions in the source message. If no location is mentioned, the report is associated to a default location in the center of Nairobi so that it may still be appear on the Uchaguzi website's map.

Finally, a separate team of volunteers is responsible for reviewing the annotated reports, ensuring the correct metadata were added, and for publishing them to the Uchaguzi website (Ushahidi, 2022c). Figure 1 shows an example of a fully annotated report.

2.2 Preprocessing

The 14,169 reports in the Uchaguzi-2022 dataset represent the subset of all incoming reports reviewed and annotated by volunteers. An additional 86k reports were received by the platform, but were left unannotated due to a lack of resources, demonstrating the need for an automated approach to address the scale of the problem. To create the annotated dataset, we exclude reports without an assigned topic (86k reports). We focus on text-based

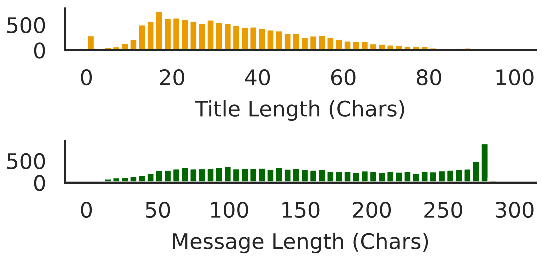


Figure 2: **Distributions of title and message lengths.**

Field	% Non-empty
`title`	97.9
`topic`	100.0
`tags`	85.1
`location`	2.4
`coordinates`	98.4

Table 3: **Non-empty fields in Uchaguzi-2022.**

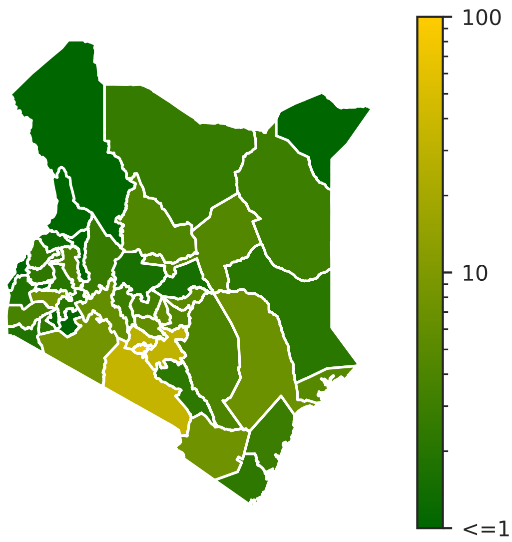


Figure 3: **Reports per capita.** Scale is per 100K citizens.

reports submitted via SMS, X (formerly Twitter), and public social media posts (as they share similar formats), and further exclude around 1,600 questionnaires (since metadata is added deterministically for this type of report). The annotated dataset along with the code are available at: `https://github.ushahidi.org/uchaguzi-ai/`.

2.3 Analysis

In this section, we include analyses of annotation coverage, lexical content, geographic distribution, and temporal trends within Uchaguzi-2022.

Annotation Coverage Each report in Uchaguzi-2022 has an assigned topic, but not all reports were further annotated with title, tags, or geotagging information (Table 3). In particular, while *coordinates* are provided for the majority of reports (98%), there is relatively low coverage of the *location* text field (2.4%). Although the reports may not mention any location, in §4.2 we find that this low coverage is also in part due to annotation incompleteness. The distribution of topic labels is shown in Table 1. The most frequent topic, *Opinions*, was assigned to 9,441 reports (66.6%), whereas the least frequent topic, *Irrelevant Report*, was assigned to 61 reports (0.4%) and generally indicated irrelevant SMS messages received by the platform.

Lexical Content In Figure 2 we present the report title and message length distributions. The spike in message lengths around 280 characters likely corresponds to the (then) Twitter length limit. Additionally, as many languages are spoken in Kenya (Muaka, 2011), we study the language distribution in the reports with `fasttext` (Joulin et al., 2017). According to this model, 98.0% of reports are in English, 1.3% are in Swahili, and the remaining 0.7% are other languages. Among the reports detected as English, we observe the presence of English-Swahili code-switching and show examples in Table 4.

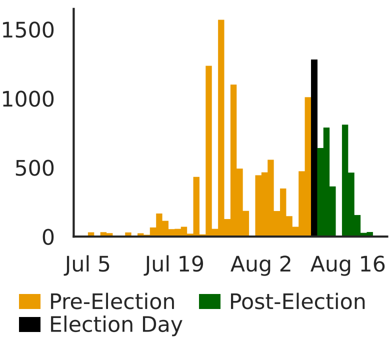
Geographic Distribution In Figure 3 we plot the per capita report count for each county in Kenya. Population sizes are based on the 2019 Kenyan Census (KNBS, 2019). Each report is associated to the county containing its *coordinates* and reports using the default location coordinates (§2.1) are omitted. We observe that the vast majority of the data comes from Nairobi and its surrounding counties.

Temporal Trends In Figure 4 we display a histogram of report counts binned by time and colored by election phase (pre-election, election day, post-election), alongside the top 10 tokens characterizing each phase. The characteristic token lists are determined by identifying the tokens that maximize the probability $P(\text{phase}|\text{token})$ estimated using a Naïve Bayes classifier over word frequencies.

The characteristic tokens provide insight into the key topics of discussion in the different phases of the election. For instance, the characteristic pre-election tokens largely correspond to a scandal where three Venezuelan employees of the firm Smartmatic were arrested at JKIA airport for possessing stolen election materials (Okoth, 2022)

Text	Translation
So why are we going to the ballot if it's so obvious to you that your candidate has already won? Anyway sisi tuna-jua kura ni 9/8/22, wa Kenya ndio wataamua na IEBC ndio watatangaza mshindi sio ng'ombe za AZIMIO.	So why are we going to the ballot if it's so obvious to you that your candidate has already won? Anyway, we know the vote is 9/8/22, Kenyans will decide and IEBC will announce the winner, not the cows of AZIMIO.
Cherera alikua chief of staff wa Joho. waliwekwa kwa commission ile time ya handshake when there was a debate about IEBC commissioners Quorum. She is out to pay her dues #kenyaelections2022 Serena kisumu Ledama	Cherera became Joho's chief of staff. they were appointed to the commission at the time of the handshake when there was a debate about IEBC commissioners quorum. She is out to pay her dues #kenyaelections2022 Serena Kisumu Ledama
DCI na IEBC nao ni kama wanatupiga kipindi, tutaona mambo before election day.	DCI and IEBC are like they are giving us a show, we will see things before election day.

Table 4: **Examples of English-Swahili code-switching in reports along with English translation**.



Pre-Election	Election Day	Post-Election
PresidentialDebate2022	RutoThe5THPresident	ElectionsKE2022
Venezuela	56.17	kenyaelections2022
JKIA	12,065,803	President-elect
Smartmatic	KIMS	138
UNDPKenya	Primary	0.01
UchaguziWaAmani	154	Serena
LetPeaceWin	Rural	34C
DCI	Rongai	uploaded
stickers	technicians	announcement
Venezuelans	violating	KenyanElection2022

Figure 4: **Report counts over time** *(left)* **and words characterizing different phases of the election** *(right)*.

(this event also corresponds to the spike in reports in late July). In contrast, the characteristic election day tokens appear to relate to preliminary results and turnout, along with reports of violations, and the characteristic post-election tokens pertain mostly to anticipation of the election results.

2.4 Data Quality

We estimate the quality of the annotations provided by the volunteers by sampling 500 reports to be further labeled by an expert annotator (also employed by Ushahidi). The 500 reports were sampled using the following strategy. From each of the topics with applicable tags (*Opinions*, *Positive Events*, *Counting and Results*, *Security Issues*, *Voting Issues*, *Polling Station Administration*), we sampled either 100 or 10% of the reports (whichever was smaller). From the remaining topics, we sampled uniformly to reach a total of 500 samples. This strategy was implemented to take into account the long-tailed nature of the topic distribution (Table 1) while ensuring a minimum number of samples for the least frequent topics. The expert annotator labeled each report with a topic and any applicable tags.

For topic, the results show 50.4% agreement between the volunteer and expert annotations (252 samples). This corresponds to a Cohen's kappa of 0.425, indicating moderate reliability (McHugh, 2012). For tags, considering the set of 252 samples where the expert and the volunteers agreed, the agreement on at least one annotated tag is 71.0%, reflected by a Cohen's kappa of 0.624. This indicates good tag reliability once the same topic has been selected.

Error analysis shows that around half of the samples annotated by the expert with a different topic were assigned to *Opinions* (in particular from *Positive Events* and *Counting and Results*). This disagreement tends to occur in cases where citizens shared their opinions on events or alleged irregularities rather than reporting factual information. An example is shown in Figure 5. In this case the citizen did not share any factual information on counting irregularities, and therefore this report should be classified as *Opinions* (expert annotation) rather than *Counting and Results* (volunteer annotation).

The overall moderate inter-annotator reliability indicates a degree of subjectivity in the dataset, par-

Title: Confirmation of results by Chebukati as seen by a voter.
Text: Chebukati knows when it comes to numbers, @RailaOdinga will win, that's why he's laying ground for nullification of Presidential elections, by messing with the process. Too bad.
Volunteer annotation: Topic: Counting and Results. Tag: Counting Irregularities.
Expert annotation: Topic: Opinions.
Tags: Personal Opinion, Negative Opinions.

Figure 5: **Example of disagreement between volunteer and expert annotations**.

ticularly for the topic label, which we believe is somewhat expected given the dataset size as well as the nature of the annotations provided by the volunteers. The comparison between the volunteer-provided and expert-provided annotations (used as "gold standard") allows us to establish a human baseline for these tasks, with disagreements reflecting the subjectivity according to non-experts. This subjectivity further supports the need for an automated approach, as this would help improve consistency in the topic and tag classifications.

3 Methods

We explore the application of Uchaguzi-2022 towards training systems for automated report *categorization* (§3.1) and *geotagging* (§3.2).

3.1 Automating Categorization

Each report is assigned a topic (see Table 1) and optionally annotated with topic-dependent tags (e.g., Table 2). Accordingly, topic prediction is a single-class classification task, while tag prediction is a multi-class classification task. We train a topic prediction model and six tag prediction models, one for each of the following topics: *Counting and Results*, *Opinions*, *Polling Station Administration*, *Positive Events*, *Security Issues*, *Voting Issues*. The remaining topics lack tags so no tag prediction model is trained. Additionally, for each task we omit labels with < 20 observations.

For each task we explore two settings: 1) fully supervised learning (FS), and 2) few-shot in-context learning (ICL). Fully supervised learning allows us to develop categorization models specific to the Uchaguzi platform. On the other hand, few-shot learning is important to explore from a practical perspective since it allows for quick adaptation of the platform in the case where topics and tags shift across elections and domains.

In the fully supervised setting, we randomly split the volunteer-labeled data into train (80%) and validation (10%) sets, and use the expert-labeled data as our test set (which was sub-sampled from the remaining 10%). In the few-shot setting, the training examples are sampled from the training split, and the test split is the same one used in the fully supervised setting. We train independent models for topic prediction and each tag prediction task using the `transformers` library (Wolf et al., 2020). We fine-tune `XLM-RoBERTa-base` (Conneau et al., 2020) for 100 epochs and select the best checkpoint using validation set performance (hyperparameters reported in Appendix B).

In the few-shot setting, we perform 25-shot learning with `mixtral-8x7b` (Jiang et al., 2024), `llama-3.1-70b` (Dubey et al., 2024), and `gpt-4o` (OpenAI et al., 2024). The LLMs chosen represent those widely used at the time of experimentation, and based on the available analyses, have some Swahili proficiency in the few-shot setting (Ochieng et al., 2024; OpenAI et al., 2024). We experiment with varying the number of examples used for in-context learning as well as with including topic and tag descriptions in the prompts. The final prompts are selected based on performance on the validation set and are shown in Appendix B.

3.2 Automating Geotagging

We decompose geotagging into two sub-tasks: 1) *location extraction*, extracting mentioned locations from reports, and 2) *geocoding*, retrieving their corresponding point coordinates.

Title: IEBC Officials Deliver Results In **Saboti** Constituency
Text: IEBC officials deliver results from various polling stations in **Saboti** constituency to tallying centre at **St Joseph's Girls High school**. #KenyaDecides2022
Location: Saboti, St Joseph's Girls High school

Figure 6: **Location extraction**. We retrieve all location mentions (**green**) in the title or text of a report.

Location Extraction Given a report title and text, we detect all mentioned locations (Figure 6). We explore NER and few-shot ICL approaches for this sub-task. As is the case for the categorization tasks, the decision to use ICL for location extraction is motivated by the adaptability of this LLM-based approach, along with the fact that the parametric knowledge of these LLMs encapsulates many Kenyan locations. Additionally, we consider only reports for which we could manually verify that the

Model	Micro F_1	Macro F_1
Human annotation (baseline)	50.4	48.6
XLM-RoBERTa-base (FS)	49.4	27.9
gpt-4o (ICL)	47.8	36.2
mixtral-8x7b (ICL)	45.0	29.3
llama-3.1-70b (ICL)	**54.6**	**39.7**

Table 5: **Topic prediction results.**

Topic	XLM-RoBERTa (FS)	llama-3.1-70b (ICL)
Opinions	63.7	**69.4**
Media Reports	51.1	**51.6**
Count. and Res.	50.6	**58.0**
Positive Events	43.7	**44.0**
Security Issues	41.7	**57.1**
Voting Issues	**28.6**	26.7
Irrelevant Report	0.0	**64.0**
Political Rallies	0.0	**26.7**
Polling Admin.	0.0	0.0
Staffing Issues	0.0	0.0

Table 6: F_1 **breakdown for the** XLM-RoBERTa (FS) **and** llama-3.1-70b (ICL) **topic prediction models.**

Title: Life beyond elections
Text: After elections we want to move with our life as if nothing happened.
Topic: Opinions

XLM-RoBERTa-base:	Positive Events
gpt-4o:	Opinions
mixtral-8x7b:	Positive Events
llama-3.1-70b:	Positive Events

Title: No voting going on at Hospital Hill Primary
Text: @IEBCKenya 6:34AM No voting going on at Hospital Hill Primary. No voting material & no IEBC officials.
Topic: Polling Station Administration

XLM-RoBERTa-base:	Security Issues
gpt-4o:	Voting Issues
mixtral-8x7b:	Voting Issues
llama-3.1-70b:	Voting Issues

Figure 7: **Examples of inaccurate topic prediction**.

annotated location appears in their title or text (147 samples). Due to this size, fine-tuning a location extraction model was unfeasible. We experiment with wikiNEuRal (Tedeschi et al., 2021) for NER, while the ICL approach relies on the same models as in §3.1.

Geocoding The extracted location names are used to query corresponding point coordinates from OpenStreetMap (OSM) using the Nominatim API[4]. We evaluate geocoding on a total of 2,818 samples: the 147 samples used for the location extraction evaluation, and, additionally, a silver dataset of 2,681 reports that do not have annotated locations but have associated coordinates and likely include a location mention. This latter set allows us to understand the extent to which location extraction systems can recover location names for reports with incomplete annotations. To obtain these evaluation sets, we begin with all reports with coordinates that do not correspond to a *default location* (§2.1). We use zero-shot prompting with gpt-4o (Appendix C) to identify and filter out reports that do not mention locations. We choose gpt-4o as it has the highest recall for location extraction (§4.2.1). We employ this approach because we observe a number of reports with labeled location but no location mentions in their title or text, making them unsuitable for geocoding.

4 Evaluation

4.1 Automating Categorization

Performance on the topic prediction and tag prediction tasks is measured on the test set using micro-averaged and macro-averaged F_1 scores.

4.1.1 Topic Prediction

The results for all models as well as the volunteer annotations (human baseline) are presented in Table 5. We observe that, overall, the ICL models are competitive with the FS model, with llama-3.1-70b outperforming all other models across both metrics. Table 6 breaks down the F_1 score by class for XLM-RoBERTa (FS) and llama-3.1-70b (ICL). While the FS model performs similarly on classes with many training examples (hence the comparable micro-averaged F_1), the ICL model outperforms on classes with fewer training examples (e.g., *Irrelevant Report* and *Political Rallies*), thus leading to a higher macro-averaged F_1. Both models fail to predict the *Polling Station Administration* and *Staffing Issues* topics. This is due to the fact that these topics have few training examples in the dataset (Table 1) and are hard to distinguish from semantically related topics, such as *Security Issues* and *Voting Issues*.

In Figure 7 we present two examples of incorrect model predictions. The first example represents the most common type of mistake across all models. Most models focus on the positive nature of this message, and do not recognize that it does not relate to any factual event (and therefore should be classified as *Opinions*). In the second example, most models identify that the citizen is reporting

[4] https://nominatim.org/

on issues pertaining to the voting process, but fail to understand that the report is about the operations of a polling station and therefore should be classified as *Polling Station Administration*. This type of mistake is more subtle and typically involves related topics (such as *Polling Station Administration*, *Voting Issues*, and *Security Issues*).

Finally, the results in Table 5 also show that the performance of `llama-3.1-70b` is competitive with the human baseline, with lower macro-averaged F_1 but higher overall accuracy. This suggests that this model could be used to create silver annotations on the unlabeled data, and aid volunteers in assigning topics.

4.1.2 Tag Prediction

Table 7 shows the tag prediction results. Overall, we observe fairly comparable performance for the different models. When larger training sets are available (e.g., *Opinions*), the FS model outperforms ICL models. In contrast, few-shot ICL performs better on the tasks with fewer training examples, with `llama-3.1-70b` as the best performing model on two tasks (*Counting and Results*, *Security Issues*). Despite the limited dataset size, all models achieve quite good performance on the *Polling Station Administration* task, with `mixtral-8x7b` obtaining a micro-averaged F_1 score of 87.0.

In Figure 8 we show two examples of incorrect model predictions. In the first example (belonging to *Voting Issues*), while the FS model does not predict any tags, the ICL models correctly predict that the report pertains to voting irregularities. However, they fail to recognize the more subtle nature of the issue, related to voting kits rather than registration procedures. In the second example (belonging to *Counting and Results*), only `llama-3.1-70b` is able to identify that the text implies counting irregularities. These mistakes occur when models cannot discriminate between semantically-related tags and represent the most common type of mistakes.

Title: KIEMS Kit Failure
Text: Limbine Primary School polling centre has 3 polling stations; and polling station 3 has been closed from 1:00pm due to kit failure. What are the voters going to do by 6:00pm? IEBC should do something.
Topic: Voting Issues
Tag: Voter Integrity Irregularities

`XLM-RoBERTa-base:`	–
`gpt-4o:`	Voter Registration Irregularities
`mixtral-8x7b:`	Voter Assistance Irregularities
`llama-3.1-70b:`	Voter Registration Irregularities

Title: Confusion over IEBC results and projections
Text: Forms delivered are over 88% but the information being translated in votes is still less than 20%.
Topic: Counting and Results
Tag: Counting Irregularities

`XLM-RoBERTa-base:`	Provisional Citizen Results
`gpt-4o:`	Failure to announce results by IEBC officials
`mixtral-8x7b:`	Failure to announce results by IEBC officials
`llama-3.1-70b:`	Counting Irregularities

Figure 8: **Examples of inaccurate tag prediction**. Minor edits to the original texts for readability.

4.2 Automated Geotagging

4.2.1 Location Extraction

Location extraction is evaluated along two dimensions: 1) the percentage of samples for which no location is extracted, and 2) how similar the extracted locations are to the annotations using text similarity metrics, such as exact match and ROUGE-L[5] (Lin, 2004). Table 8 shows larger models achieve the highest performance in both exact match and ROUGE-L, with `llama-3.1-70b` outperforming the other models. We observe a large gap in performance between the mBERT-based `wikiNEuRal` model and the ICL approaches. Although `wikiNEuRal` is generally able to identify mentions of large cities (e.g., Nairobi, Mombasa), it performs poorly at detecting specific landmarks, hence the overall very low performance. As `WikiNEuRal` was trained on 50k English Wikipedia articles, it is unlikely that the training data covered the specific locations present in our evaluation set.

4.2.2 Geocoding

We report three metrics commonly used to evaluate geocoding: coverage (percentage of queries that return coordinates), accuracy at 161 km (Acc@161km), and area under the curve (AUC). Acc@161km computes the percentage of locations which are correctly predicted to within 161 km (100 miles) of the actual location. On the other hand, AUC, computed using the trapezoidal rule (Yeh et al., 2002), measures the error across all instances while minimizing the importance of outliers. AUC for geocoding was introduced by Jurgens et al. (2015) and it allows for relative comparison of geocoders with similar Acc@161km scores.

We present results in Table 9 (samples with labeled location) and Table 10 (samples without la-

[5]ROUGE-L is chosen to fairly evaluate generative approaches, as they may produce location names that resemble but do not exactly match spans in the source text.

Task	XLM-RoBERTa (FS)	gpt-4o (ICL)	mixtral-8x7b (ICL)	llama-3.1-70b (ICL)
Counting and Res.	38.4	49.5	44.9	**52.3**
Opinions	**51.0**	49.5	35.4	38.2
Polling Admin.	77.8	76.2	**87.0**	81.8
Positive Events	55.4	**65.7**	56.7	59.2
Security Issues	58.3	64.3	73.3	**74.1**
Voting Issues	**60.8**	38.7	48.3	38.7

Table 7: **Tag prediction results.** Performance is measured using micro-averaged F_1.

Model	% Empty Pred. (↓)	Exact Match (↑)	ROUGE-L (↑)
wikiNEuRal	20.4	1.4	14.2
gpt-4o	**3.4**	46.3	71.5
mixtral-8x7b	10.2	40.1	67.6
llama-3.1-70b	4.8	**49.0**	**75.3**

Table 8: **Location extraction results for 147 samples with labeled location.** Exact match and ROUGE-L are averaged across extracted locations.

Model	Cov. (↑)	Acc@ 161km (↑)	AUC (↓)
oracle	53.7	43.5	337
wikiNEuRal	79.6	6.1	495
gpt-4o (ICL)	86.4	66.7	156
mixtral-8x7b (ICL)	78.2	62.6	209
llama-3.1-70b (ICL)	**87.1**	**68.7**	**123**

Table 9: **Geocoding results for 147 samples with labeled location.** AUC is expressed in 1,000s of km^2.

Model	Cov. (↑)	Acc@ 161km (↑)	AUC (↓)
wikiNEuRal	36.3	1.9	9,218
gpt-4o (ICL)	**80.6**	38.2	4,245
mixtral-8x7b (ICL)	63.6	**41.1**	5,540
llama-3.1-70b (ICL)	80.2	38.2	**3,674**

Table 10: **Geocoding results for 2,681 samples without labeled location.** AUC is expressed in 1,000s of km^2.

Title: Civic Education Campaign
Text: Day 2 of our Civic Education campaign we are in **Changamwe Social Hall** having a dialogue with young people on why the should vote and the do's and dont's during the voting day with @siasaplace and @IEBC_YCC #GE2022
Coordinates: $(-4.02084, 39.62748)$

gpt-4o:	Changamwe Social Hall
Retrieved Coordinates:	–

Figure 9: **Example of failed geocoding**.

beled location). Where no coordinates are returned, we use Null Island[6] as the reference to compute the error. For the oracle in Table 9, the annotated location names serve as input queries for geocoding.

For samples with labeled location (Table 9), we observe that while all location extraction models have good coverage, the ICL approaches clearly outperform the NER approach on accuracy and AUC. In particular, the higher location extraction performance observed in Table 8 for llama3.1-70b provides this model with an advantage over all other approaches for the downstream task. Across approaches, geocoding errors most frequently tend to occur for specific building names. The oracle performs poorly compared to the ICL approaches because the annotated locations can be more specific and therefore less likely to be resolved by the OSM query than the less precise locations extracted by the ICL models. Finally, we observe that although the locations extracted using wikiNEuRal achieve high coverage, the returned coordinates are mostly inaccurate.

The results for samples without labeled location (Table 10) show comparable performance among the ICL models. Comparing the results for samples with labeled locations and those without, we observe a marked decrease in performance across all metrics, most starkly for Acc@161km and AUC. However, the results show that the ICL methods are still able to correctly geocode around 40% of reports without explicit location annotations, providing evidence that a lack of annotated location does not necessarily mean that the report does not mention any. Our geocoding approaches tend to fail for two reasons: either the queried location is not present in the OSM database, or the location is resolved but OSM returns coordinates that are vastly different from those in the labeled data, resulting in a significant error. We believe better map coverage, particularly for building names and landmarks, could further improve performance.

Figure 9 illustrates such an example. In this case the report was geotagged with coordinates but no explicit location was added. While gpt-4o is

[6] https://en.wikipedia.org/wiki/Null_Island

able to extract the correct location mention, the exact building name cannot be resolved into point coordinates using OpenStreetMap and therefore `gpt-4o` fails to geocode the report.

5 Related Work

Alongside the rise of social media, election analysis has become a popular topic in NLP. Studies in this field have typically focused on tasks such as electoral sentiment analysis (Wang et al., 2012; Schmidt et al., 2022; Hellwig et al., 2023), predicting election outcomes (Bermingham and Smeaton, 2011; Tjong Kim Sang and Bos, 2012; Sanders and van den Bosch, 2020), detecting online political communities (Wan and Paris, 2015; Abdine et al., 2022), electioneering (Baran et al., 2022), predicting success rates of campaigning strategies (Litvak et al., 2022; Mohapatra and Mohapatra, 2022; Bhaumik et al., 2023), and analyzing election fraud claims (Abilov et al., 2021). Our work differs from these studies in a number of ways.

Firstly, existing literature mostly centers on elections in the United States or Europe, whereas Uchaguzi-2022 covers Kenyan elections and politics. While there are other works that focus on the 2022 elections (Amol et al., 2024), their emphasis is on creating a political misinformation dataset. Secondly, Uchaguzi-2022 is novel in its focus on detecting and categorizing reports of electoral misconduct, voting issues, and related acts of violence. Although there exist other datasets studying issues of civil unrest, Chinta et al. (2021) does not include fine-grained topic and geographical information, while the datasets introduced in Raleigh et al. (2010) and Sundberg and Melander (2013) are based on news articles and structured reports from NGOs rather than citizen reports. Uchaguzi-2022 is unique among these datasets as it also links reported events to their locations. Lastly, in addition to being a useful resource for training models to identify reports of electoral misconduct, we believe Uchaguzi-2022 could also be of value to the topic of geotagging noisy social media posts (Seeberger and Riedhammer, 2022; Li and Lim, 2024), as the majority of reports are X (formerly Twitter) posts.

6 Conclusion

In this paper we introduced Uchaguzi-2022, a dataset of 14k citizens' opinions and reports on the 2022 Kenyan General Election, including mentions of electoral misconduct and election-related acts of violence. We demonstrated potential applications of this dataset by training and benchmarking language models for automated report categorization and geotagging, achieving viable performance while leaving ample room for further improvement. Initial explorations integrating these models into a live system indicate that the performance of the topic and tag prediction models can improve efficiency in generating insights in rapidly evolving situations (Ushahidi, 2024), while next steps are needed to evaluate the geotagging models in this setting. We additionally presented preliminary analyses of the presence of different languages, code-switching, and geo-spatial and temporal properties of the reports in this dataset. We believe these demonstrate the dataset's potential for deeper sociolinguistic analyses. Finally, we hope this dataset will be a valuable resource for studying NLP applications to election integrity monitoring in areas of the world where such systems could be crucial to help ensure fair elections, thus further advancing AI for Social Good efforts.

7 Limitations

The topic labels used to annotate the dataset are not specific to the context of Kenyan elections, and therefore we believe that our approach for the topic prediction task can be applied to elections in other geographical regions. On the other hand, the topic-dependent tags are specific to Kenya (e.g., they reference the Independent Electoral and Boundaries Commission). Tasks focusing on elections in other countries would require a new set of applicable tags to be defined.

We acknowledge that for some tag prediction tasks test set sizes are limited, restricting our ability to compare the performance of different models. We believe this further highlights the need for datasets such as Uchaguzi-2022, as manually processing large amount of reports to collect a sufficient number of examples for each topic and tag category is intractable.

8 Ethical Considerations

We would like to emphasize that the dataset presented in this work contains only data that is already publicly available online and collected in accordance with the terms of service of the Uchaguzi platform. None of the reports are confidential, and care has been taken to remove any personally identifiable information (such as names and phone numbers). Furthermore, the geographic coordinates associated to the reports reflect the physical location of the reported event and *not* the location of the citizen reporting the event.

In order to access and download the full dataset, researchers will need to fill out a licensing agreement prohibiting them from using the data for malicious purposes. Accordingly, we do not anticipate the publication of this dataset would endanger any person or persons.

Acknowledgments

This work was made possible by Dataminr's AI for Good program and we are grateful to Sirene Abou-Chakra and Jessie End for their support during this project. We thank the partners at Ushahidi for their support and collaboration. We also wish to thank the reviewers and meta-reviewer for their valuable feedback.

References

Per Sekse Aarvik. 2015. Uchaguzi: An analysis of the crowdsourced election monitoring in kenya 2013. Master's thesis, The University of Bergen.

Hadi Abdine, Yanzhu Guo, Virgile Rennard, and Michalis Vazirgiannis. 2022. Political communities on Twitter: Case study of the 2022 French presidential election. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 62–71, Marseille, France. European Language Resources Association.

Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 901–912. AAAI Press.

Khadijat Oluwatoyin Ajao. 2017. *Citizen journalism and conflict in Africa : the Ushahidi Platform in Kenya's 2008 post-election violence*. Thesis, University of Pretoria. Accepted: 2018-07-16T07:53:39Z.

Cynthia Amol, Lilian Wanzare, and James Obuhuma. 2024. Politikweli: A swahili-english code-switched twitter political misinformation classification dataset. In *Speech and Language Technologies for Low-Resource Languages*, pages 3–17, Cham. Springer Nature Switzerland.

Mateusz Baran, Mateusz Wójcik, Piotr Kolebski, Michał Bernaczyk, Krzysztof Rajda, Lukasz Augustyniak, and Tomasz Kajdanowicz. 2022. Electoral agitation dataset: The use case of the Polish election. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 32–36, Marseille, France. European Language Resources Association.

Adam Bermingham and Alan Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pages 2–10, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ankita Bhaumik, Andy Bernhardt, Gregorios Katsios, Ning Sa, and Tomek Strzalkowski. 2023. Adapting emotion detection to analyze influence campaigns on social media. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 441–451, Toronto, Canada. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Abhinav Chinta, Jingyu Zhang, Alexandra DeLucia, Mark Dredze, and Anna L. Buczak. 2021. Study of manifestation of civil unrest on Twitter. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 396–409, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Nils Constantin Hellwig, Markus Bink, Thomas Schmidt, Jakob Fehle, and Christian Wolff. 2023. Transformer-based analysis of sentiment towards German political parties on Twitter during the 2021 election year. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 84–98, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 188–197. AAAI Press.

KNBS. 2019. 2019 Kenya population and housing census. Technical report, Kenya National Bureau of Statistics.

Menglin Li and Kwan Hui Lim. 2024. Leveraging contrastive learning for few-shot geolocation of social posts. *ArXiv preprint*, abs/2403.00786.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Marina Litvak, Natalia Vanetik, Sagiv Talker, and Or Machlouf. 2022. Detection of negative campaign in israeli municipal elections. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 68–74, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Sovesh Mohapatra and Somesh Mohapatra. 2022. Sentiment is all you need to win US presidential elections. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 15–20, Taipei, Taiwan. Association for Computational Linguistics.

Dumisani Moyo. 2009. Citizen Journalism and the Parallel Market of Information in Zimbabwe's 2008 Election. *Journalism Studies*, 10(4):551–567. Publisher: Routledge _eprint: https://doi.org/10.1080/14616700902797291.

Leonard Muaka. 2011. Language perceptions and identityamong kenyan speakers. In *Proceedings of the 40th Annual Conference on African Linguistics*.

Martin N Ndlela. 2020. Social media algorithms, bots and elections in africa. *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns*, pages 13–37.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond metrics: Evaluating llms' effectiveness in culturally nuanced, low-resource real-world scenarios. *CoRR*, abs/2406.00343.

Brian Okoth. 2022. Chebukati: Three Venezuelans key in Aug. 9 polls technology roll-out arrested at JKIA. *The Standard*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Clionadh Raleigh, rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Eric Sanders and Antal van den Bosch. 2020. Optimising Twitter-based political election prediction with relevance andSentiment filters. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6158–6165, Marseille, France. European Language Resources Association.

Thomas Schmidt, Jakob Fehle, Maximilian Weissenbacher, Jonathan Richter, Philipp Gottschalk, and Christian Wolff. 2022. Sentiment analysis on Twitter for the major German parties during the 2021 German federal election. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 74–87, Potsdam, Germany. KONVENS 2022 Organizers.

Philipp Seeberger and Korbinian Riedhammer. 2022. Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 70–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Deodatus Patrick Shayo. 2021. Citizen participation in local government elections in the age of crowdsourcing: Explorations and considerations in tanzania. *Program on Governance and Local Development Working Paper*, (47).

Ralph Sundberg and Erik Melander. 2013. Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Erik Tjong Kim Sang and Johan Bos. 2012. Predicting the 2011 Dutch senate election results with Twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 53–60, Avignon, France. Association for Computational Linguistics.

Ushahidi. 2022a. Uchaguzi platform.

Ushahidi. 2022b. Ushahidi's digital response teams.

Ushahidi. 2022c. Ushahidi's publishing team.

Ushahidi. 2024. Report on sentiment around protests in kenya.

Stephen Wan and Cécile Paris. 2015. Ranking election issues through the lens of social media. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 48–52, Beijing, China. Association for Computational Linguistics.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shi-Tao Yeh et al. 2002. Using trapezoidal rule for the area under a curve calculation. *Proceedings of the 27th Annual SAS® User Group International (SUGI'02)*, pages 1–5.

A Tag Sets

We provide the complete list of applicable tags within each topic in Tables A1-A7 along with their prevalence. Reports belonging to the following topics are not further assigned any tags: *Media Reports*, *Political Rallies*, and *Irrelevant Report*. Table A8 shows the list of topics with example messages.

Tag	Count
Unofficial proclamation of results	220
Provisional citizen results	173
Counting irregularities	74
Official IEBC results	64
Protest over declared results	43
Failure to announce results by IEBC official	33
Fake 34 result forms	17
Party agent irregularities	5
Irregularities with transportation of ballot boxes	3

Table A1: **Counting and Results**.

Tag	Count
Personal opinion	3,577
Positive opinions	2,392
Negative opinions	1,887
Neutral	1,543
Peace messages	435

Table A2: **Opinions**.

Tag	Count
Biometric voter registration (BVR) issues	38
Polling station logistical issues	38
Low voter turn out	23
Missing/inadequate voting materials	14
Ballot box irregularities	9
Polling station closed before voting concluded	5
Campaign material in polling station	3
High voter turn out	2

Table A3: **Polling Station Administration**.

Tag	Count
Citizen led initiatives to promote peace	925
Everything fine	306
Organization led initiatives	58
Police peace efforts	18
Civic education	5

Table A4: **Positive Events**.

Tag	Count
Rumors	117
Dangerous speech	62
Violent attacks	32
Abductions/kidnapping	26
Mobilisation towards violence	26
Demonstrations	24
Heavy police presence	22
Vandalism and physical attacks on property	21
Riots	15
Presence of weapons	10
Armed clashes	8
Ambush	7
Eviction/population displacement	4
Police brutality	3
Sexual and gender based violence	3

Table A5: **Security Issues**.

Tag	Count
IEBC officials not acting in accordance to set rules	65
Absence or insufficient number of IEBC officials/staff at polling station opening	8
Observers/media blocked from entering polling station	2
Absence or insufficient number of law enforcement officials at polling station	1

Table A6: **Staffing Issues**.

Tag	Count
Voting irregularities	69
Voter integrity irregularities	66
Voter registration irregularities	34
Civic education gap	33
Voter assistance irregularities	28
Alleged rigging	21
Voters issued invalid ballot papers	11
Voting suspended/postponed	10
Purchasing of voters cards	9
Voter intimidation and blockage	7

Table A7: **Voting Issues**.

Topic	Example Message
Opinions	So shameful that #wafulachebukati IEBC Chairman is once again in headlines for the same wrong reasons of 2017. In civilized countries once you have been found to lack integrity one resigns.
Media Reports	William Ruto: "We made a commitment we are going to have a Diaspora Ministry so that the many issues that concern our people in the diaspora are sorted out" #KenyaDecides
Positive Events	There was a peaceful and organised elections in bondo constituency siaya county
Counting and Results	William Ruto's losses 10,000 votes in Kiambu constituency as IEBC corrects the votes from 51,050 to 41,050.
Security Issues	Reports of road blockage on the outer ring road in NBO. Southern bypass near Kibra also seeing worsening disruption. Avoid area.
Voting Issues	Voters funding it difficult to get their names in voters register; some of those who talked to us say IEBC officials can't help them
Political Rallies	Deputy President (incoming) Rigathi Gachagua will tour Narok & Kajiado counties today. Kenya Kwanza RG's Itinerary: 1. Narok Town 2. Nairagie Enkare 3. Uwaso 4. Ilasit (KAJIADO South) 5. Loitoktok 6. Kimana IEBC Dismus
Polling Station Administration	IEBC failed to deliver manual registers to every polling station. Just seen 50 votes for Baba getting lost in my station
Staffing Issues	Eight IEBC officials from Homabay, Kisumu and Bungoma counties have been arrested and sacked after they were found meeting two candidates for parliamentary and county assembly.
Irrelevant Report	If i dont have ID i can vote?

Table A8: **List of topics and example messages**.

B Classification

For the fully-supervised setting, we finetune models using AdamW (Loshchilov and Hutter, 2019) optimizer with the hyperparameters specified in Table B1. Finetuning was performed on machines with one NVIDIA A10g GPU (24 GB of VRAM).

In the few-shot in-context learning setting, the prompt for the topic prediction task is shown in Table B2, while the prompts for the tag prediction tasks are shown in Tables B3-B8. For tag prediction, outputs are requested to be provided as JSON-formatted arrays, and are parsed by truncating the output string after the first occurrence of a] character and then passed through a JSON parser. We did not encounter formatting issues.

Learning Rate	1e-5
β_1	0.9
β_2	0.999
ϵ	1e-8
Max. Grad. Norm	1.0
Batch Size	8

Table B1: **Fine-tuning hyperparameters.**

```
You are a topic classifier. Given an input
text you will classify it with one and only
one of the following topics:
Counting and Results
Staffing Issues
Positive Events
Opinions
Political Rallies
Polling Station Administration
Irrelevant Report
Security Issues
Media Reports
Voting Issues
```

Table B2: **Prompt for topic prediction.**

```
You are a topic classifier. Given an input
text you will classify it with zero or more
of the following topics output in a JSON array
format:
Label: Failure to announce results by IEBC
officials, Description: Failure to announce
results by IEBC officials
Label: Protest over declared results,
Description: Violence and demonstrations
revolving around voting results
Label: Unofficial proclamation of results,
Description: Unofficial proclamation of
results
Label: Official IEBC results, Description:
Official IEBC results
Label: Provisional Citizen Results,
Description: Provisional Citizen Results
Label: Counting Irregularities, Description:
Involves Ballot Papers not Being Counted in
a Transparent Manner, Observers or Party
Agents not Allowed In The Hall During
Vote Counting, Spoiled Ballot Papers not
Properly Preserved For Review, Intimidation
of Counting Officials & Observers, Error or
Omission In Computing or Completing Tally
Sheets, Unusually Many Rejected/Spoilt Ballot
Papers, Officials Tallying Wrong/Tampered
Results, Officials not Reporting Results At
Prescribed Time, etc.
```

Table B3: **Prompt for *Counting and Results* tag prediction.**

```
You are a topic classifier. Given an input
text you will classify it with zero or more
of the following topics output in a JSON array
format:
Negative opinions
Neutral
Peace messages
Personal Opinion
Positive Opinions
```

Table B4: **Prompt for *Opinions* tag prediction.**

```
You are a topic classifier. Given an input
text you will classify it with zero or more
of the following topics output in a JSON array
format:
BVR issues
Low voter turn out
Polling station logistical issues
```

Table B5: **Prompt for *Polling Station Administration* tag prediction.**

```
You are a topic classifier. Given an input text you will classify it with zero or more of the following topics output in a JSON array format:
Label: Citizen led initiatives to promote peace, Description: Citizen led initiatives to promote peace
Label: Everything Fine, Description: People reporting that things are going smoothly.
Label: Organization led Initiatives, Description: The police sometimes have community outreach events to promote peace and security during the polls.
```

Table B6: **Prompt for *Positive Events* tag prediction.**

```
You are a topic classifier. Given an input text you will classify it with zero or more of the following topics output in a JSON array format:
Label: Demonstrations, Description: Rallies and Marches
Label: Mobilisation towards violence, Description: Situations where violence is imminent but has not started
Label: Violent Attacks, Description: Attacks involving unarmed combatants, knives, or few (1 to 2) handguns
Label: Abductions/kidnapping, Description: Acts of abductions, kidnapping, or hostage taking
Label: Rumors, Description: Presence of a credible threat, along with unverified reports of violence, corruption, looting, etc.
Label: Dangerous Speech, Description: Threats of violence
Label: Vandalism and Physical Attacks on Property, Description: Vandalism and Physical Attacks on Property
Label: Heavy police presence, Description: Presence of a large number of police officers
```

Table B7: **Prompt for *Security Issues* tag prediction.**

```
You are a topic classifier. Given an input text you will classify it with zero or more of the following topics output in a JSON array format:
Label: voting irregularities, Description: Cases such as eligible voters being turned away or not allowed to vote, and ineligible voters allowed to vote
Label: voter integrity irregularities, Description: Cases such as importation of voters, voter impersonation, voter intimidation, bribing of voters, voters voting more than once, voter identification kit not working, etc.
Label: voter registration irregularities, Description: Issues with registration such as the register of voters missing, or voter names missing from the registry
Label: civic education gap, Description: Voter requesting or lacking crucial information about how to vote
Label: voter assistance irregularities, Description: Cases such as issues with proper identification, illiterate voters not being assisted, unusually many assisted voters, and the voter assister not taking the oath of secrecy
Label: alleged rigging, Description: Allegations of rigging elections
```

Table B8: **Prompt for *Voting Issues* tag prediction.**

C Geotagging

Figure C1 shows the zero-shot prompt used to select the 2,818 reports for the geocoding evaluation. This prompt is used to identify reports that contain locations in either their title or text. Figure C2 shows the prompt used for the location extraction task along with the examples provided for in-context learning.

```
You are an advanced text analysis model.
Your task is to determine whether a given
title and text contain a named location
(e.g., a country such as Kenya, city such
as Mombasa, landmark such as the Masaai Mara,
or geographical region such as Nyanza). Named
locations include any proper nouns that refer
to places.
Instructions:
Input: You will be provided with a "Title"
and "Text."
Output: You should output "Yes" if you detect
the presence of a named location in either
the title or text, and "No" if there are no
named locations.
```

Figure C1: **Prompt used to predict the presence of a location mention in a report.**

D License

The dataset is available for research purposes only, upon submission and review of a data request form.

```
You are a helpful, respectful and honest assistant. You are have been given the task of named location extraction. If you don't know the answer to a question, please don't share false information. Extract the exact match for all named location in the following text. Output a JSON object with the a single location field containing a list (e.g. {"location": ["Nairobi", "Nairobi County"]}).

Example 1:
Title: "IEBC Chairperson Wafula Chebukati Calls for Prayers for Staff Families"
Text: "Wafula Chebukati, the Independent Electoral and Boundaries Commission (IEBC) chairperson in Bomas of Kenya, has appealed to Kenyans to pray for the spouses and children of his staff."
Output: {"label": ["Bomas of Kenya"]}

Example 2: Title: "Nairobi Voters Speak Out: A Look at the 2022 Kenyan General Elections" Text: "As the 2022 Kenyan general elections approach, voters in the capital are eagerly awaiting their chance to make their voices heard. With a diverse range of candidates and issues on the ballot, the city's residents are poised to play a crucial role in shaping the future of the country."
Output: "label": ["Nairobi"]

Example 3:
Title: "The Future of Agriculture: A Look at the 2022 Kenyan General Elections"
Text: "As the 2022 Kenyan general elections approach, the future of agriculture is a top concern for many voters. With issues such as land use, irrigation, and crop yields at the forefront of the campaign, we will be hosting events in Nakuru, Eldoret and Nyeri to take a closer look at some of the candidates' positions on these critical issues."
Output: "label": ["Nakuru", "Eldoret", "Nyeri", "Nakuru, Eldoret and Nyeri"]

Example 4:
Title: "Bomet Teachers Training College Students Endorse Local Candidate in 2022 Kenyan General Elections"
Text: "At Bomet Teachers Training College, Kaplong-Narok-Maai Road located in Bomet in the Rift Valley, students have come together to endorse a local candidate in the upcoming general elections. The candidate, who has been actively involved in community service, has gained the support of the students due to his commitment to improving the lives of the residents."
Output: {"label": ["Bomet Teachers Training College", "Bomet Teachers Training College, Kaplong-Narok-Maai Road", "Kaplong-Narok-Maai Road", "Bomet", "Rift Valley", "Bomet Teachers Training College, Kaplong-Narok-Maai Road located in Bomet in the Rift Valley"]}

Example 5:
Title: "Manyatta Estate Residents Demand Better Representation in Upcoming Elections"
Text: "Residents of Manyatta Estate in Kisumu, Western Kenya, are calling for better representation in the upcoming 2022 Kenyan general elections. With a growing population and a lack of basic amenities, the community is eager to have their voices heard and their needs addressed."
Output: "label": ["Manyatta Estate", "Kisumu, Western Kenya", "Kisumu", "Western Kenya"]

Example 6:
Title: "Mwingi MP Victor Munyasia Hosts Public Debate Ahead of 2022 Elections"
Text: "Mwingi MP Victor Munyasia recently hosted a public debate for residents of Mwingi and surrounding areas to discuss issues affecting the community and their hopes for the 2022 Kenyan general elections."
Output: "label": ["Mwingi"]

Example 7:
Title: "Busy Polling Station"
Text: "Hey! Just got to the Westlands Primary School polling station in Westlands Constituency. The place is packed, lines are super long but everyone seems determined. People are chatting, sharing snacks, and even singing. It's a great vibe. Voting is slow but steady. Make sure you bring water and a hat if you're coming later. #KenyaDecides2022"
Output: {"label": ["Westlands Primary School", "Westlands Primary School polling station", "Westlands", "Westlands Constituency", "Westlands Primary School polling station in Westlands Constituency"]}

New task:
Title: "Urging The Youth To Promote Peace And Liberty"
Text: At PrideInn Paradise Mombasa encouraging YOUTH to promote peace and liberate themselves as well as the society at large with an aim to prevent a repeat of the 2007 PEV #UchaguziWaAmani @YEDNetworkKe @VybezRadioKE #Vijanapeace
Output:
```

Figure C2: **Prompt for location extraction.**